Predicting the popularity of spotify songs with Byesian Networks

Benkherfallah Dounia

Master's Degree in Artificial Intelligence, University of Bologna

dounia.benkherfallah@studio.unibo.it

Abstract

This report presents a probabilistic approach to predicting song popularity using Bayesian Networks. Inspired by a similar methodology applied to video game sales prediction, we constructed and compared three Bayesian Network models: Tree-Augmented Naive Bayes (TAN), Tree Search, and Tabu Search. The models were trained on a preprocessed dataset containing features such as track_genre, danceability, loudness, tempo, and others, with the target variable popularity. Conditional Probability Distributions (CPDs) were calculated for each model, and insightful inference questions were addressed to evaluate their performance. The results provide actionable insights into how specific features and their combinations influence song popularity, offering valuable guidance for artists and music producers.

Introduction

Domain

Music popularity prediction is a critical task in the music industry, as it helps its actors understand the factors that contribute to a song's success. With the rise of digital music platforms, vast amounts of data are now available. By leveraging this data, we can build predictive models to identify the key drivers of song popularity and provide actionable recommendations for artists and industry stakeholders.

Aim

The aim of this project is to predict the popularity of songs using Bayesian Networks. We explore three distinct network structures: **Tree-Augmented Naive Bayes (TAN)**, **Tree Search**, and **Tabu Search**. By comparing these models, we aim to identify the most effective approach for understanding the relationships between song features and popularity. Additionally, we seek to answer key questions, such as which track genre maximizes popularity.

Method

We preprocessed the dataset to ensure all features were categorical, and with the use of the pgmpy library we constructed each network. Conditional Probability Distributions (CPDs) were calculated for each model using Maximum Likelihood **Elimination** algorithm to answer specific questions about the relationships between features and song popularity. Visualizations were created to compare the results across models and provide actionable insights.

Results

The high danceability and the high valence of a track don't always imply that a son gis going to be a hit. It means that calmer songs work as well . We also found that the genre of the song has an impact on its popularity and finaly some combination of features for the most popular songs didn't give us the results we were expecting. A song can't be of high danceability and have a low tempo it doesn't make much sense.

Model

A TAN is a model that extends the Naïve Bayes classifier by allowing dependencies between predictor variables , forming a tree structure to improve classification accuracy while maintaining computational efficiency.

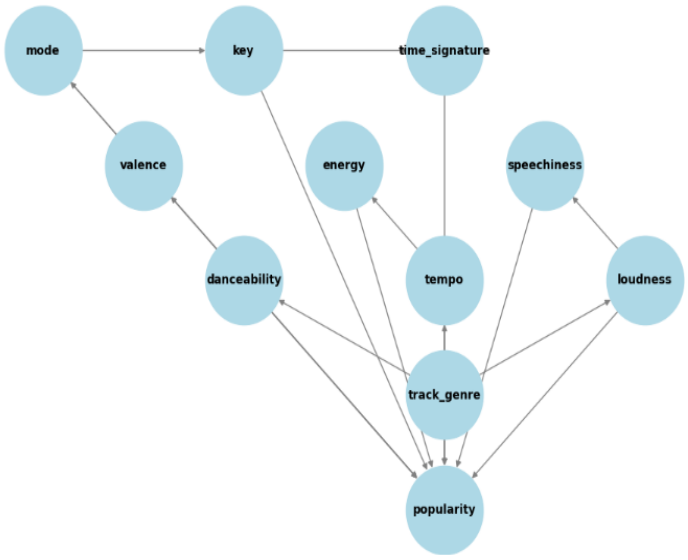

Figure 1:TAN Structure

We constructed this TAN trying to make sense of the different features and how they would influence each other in the music domain. For exemple we know the link beween the tempo and the energy of the song etc … [1]

Tree search is chosen for systematic exploration of decision spaces, by systematically expanding and evaluating nodes in a tree structure. It is widely applied in problem-solving, pathfinding,and game-playing. It ensures optimal or near-optimal solutions in structured problems.

We chose the track_genre as the root node . A well-defined root can optimize search performance by reducing unnecessary computations.
Estimation. Inference was performed using the **Variable**

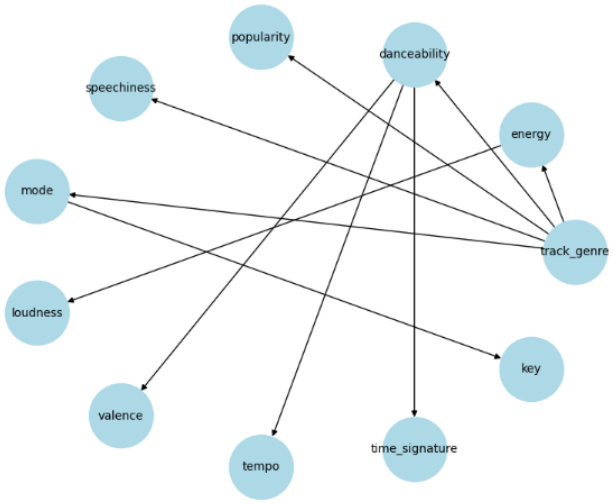

Figure 2: Tree Search Graph

Tabu Search is a metaheuristic optimization algorithm used to find the optimal or near-optimal structure of a Bayesian network. The goal is to maximize a scoring metric (in our case BDeu) that evaluates how well the network fits the data.

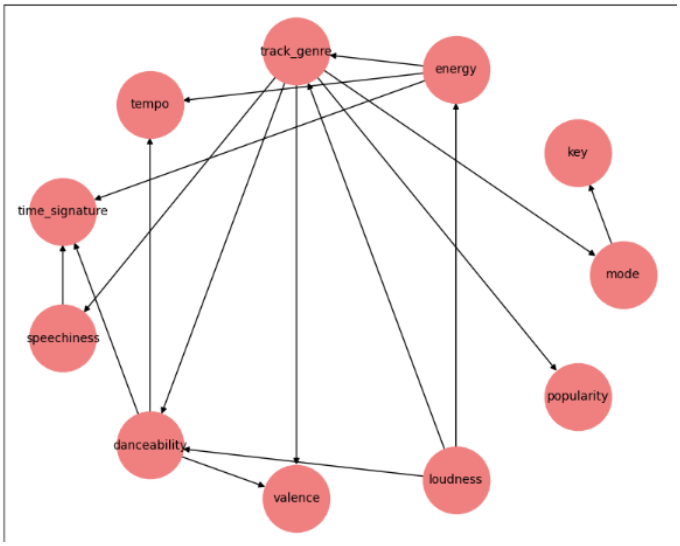

Figure 3: Tabu Search Network

Our tabu Search network contains some interesting relationships as well as unusual ones.

Analysis

Experimental setup

We carried some queries on our different networks. In the exploration of danceabilty and valence on popularity, we expected that the tracks with higher values would be most popular, since they would be played in clubs .
In the query dedicated to genres, pop, worldmusic and popular genres found trending on social media would dominate the charts.
Finally we tried to find the best combination of feture ton augment the chances of popularity of a song.

Results

Weirdly, the songs with lower valence and lower danceability were the one with the highest probability, but not with a significant difference
The genres with the most popularity are R&B, Anbiancy/Chill and world music.
The optimal Combination between Danceability=low, Loudness=low, Tempo=high, the highest probabilty observed in popularity is 0.48 .

Conclusion

In this study, we explored the use of Bayesian Networks to predict song popularity based on features such as track_genre, danceability, loudness, and tempo. By constructing and comparing three models—**TAN**, **Tree Search**, and **Tabu Search**—we identified key relationships between song features and popularity. Our results highlight that **genre selection**, **loudness**, and **tempo** play significant roles in determining a song's success, with specific combinations of these features maximizing the probability of high popularity. These insights provide actionable recommendations for artists and industry professionals, enabling them to make data-driven decisions to enhance song appeal. Future work could extend this approach by incorporating additional features or exploring other probabilistic models to further refine predictions.

Links to external resources

https://www.kaggle.com/datasets/thedevastator/spotify-tracks-genre-dataset?select=train.csv

https://github.com/dounia-bnk/Bayesan-Networks-on-Song-Popularity/tree/main

References

[1] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). *Bayesian Network Classifiers*. Machine Learning, 29(2-3), 131–163.
[2] Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.
[3] Glover, F. (1989). *Tabu Search—Part I*. ORSA Journal on Computing, 1(3), 190–206.